Machine Learning Theory (CS 6783)

Lecture 14: Online Mirror Descent

1 Recap

$\mathcal{F}$ is a convex subset of a vector space.

For $t = 1$ to n

Learner picks $\hat{\mathbf{y}}_t \in \mathcal{F}$

Receives instance $z_t \in \mathcal{Z}$

Suffers convex loss $\ell(\hat{\mathbf{y}}_t, z_t)$

End

The goal is to minimize regret :

$$\mathrm{Reg}_n := \frac{1}{n}\sum_{t=1}^{n} \ell(\hat{\mathbf{y}}_t, z_t) - \inf_{\mathbf{f}\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(\mathbf{f}, z_t)$$

It suffices to be able to solve online linear optimization. This is because, for any $f \in \mathcal{F}$, by convexity, $\ell(\hat{\mathbf{y}}_t, z_t) - \ell(\mathbf{f}, z_t) \le \langle \nabla \ell(\hat{\mathbf{y}}_t, z_t), \hat{\mathbf{y}}_t - f\rangle$, and so,

$$\mathrm{Reg}_n \le \frac{1}{n}\sum_{t=1}^{n} \langle \nabla_t, \hat{\mathbf{y}}_t\rangle - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \langle \nabla_t, \mathbf{f}\rangle$$

- **Online Gradient Descent Algorithm :**

$$\hat{\mathbf{y}}_{t+1} = \Pi_{\mathcal{F}}\left(\hat{\mathbf{y}}_t - \eta \nabla_t\right)$$

- $\eta = \frac{R}{B\sqrt{n}}$ and $\hat{\mathbf{y}}_1 = \mathbf{0}$, then $\mathrm{Reg}_n \le \frac{RB}{\sqrt{n}}$ where $\sup_{\mathbf{f}\in\mathcal{F}} \|\mathbf{f}\|_2 \le R$ and $\sup_{\nabla\in\mathcal{D}} \|\nabla\|_2 \le B$

2 Online Mirror Descent

Is the online gradient descent algorithm always the right thing to use? Let us look at the finite experts problem. $\mathcal{F} = \Delta_N$ and $\langle \mathbf{f}, \nabla_t\rangle = \mathbb{E}_{i\sim\mathbf{f}}\left[\phi(i, z_t)\right]$. Notice that in this setting, for any $\mathbf{f} \in \Delta_N$, $\|\mathbf{f}\|_2 \le \|\mathbf{f}\|_1 = 1$. However note that $\|\nabla_t\|_2 = \sqrt{\sum_{i=1}^{N} |\phi(i, z_t)|} \le \sqrt{N}$ (assuming losses are bounded by 1). Hence GD bound can only given a rate of

$$\mathrm{Reg}_n \le \sqrt{\frac{N}{n}}$$

But is this the best rate possible? In statistical learning setting we know that $\log N$ was achievable, can we obtain that here? What is the right algorithm in general. In fact in general vector spaces, GD does not even type check!

Strongly convex function: Function R is said to be λ-strongly convex w.r.t. norm $\|\cdot\|$ if $\forall \mathbf{f}, \mathbf{f}'$,

$$R\left(\frac{\mathbf{f}+\mathbf{f}'}{2}\right) \leq \frac{R(\mathbf{f})+R(\mathbf{f}')}{2} - \frac{\lambda}{2}\left\|\mathbf{f}-\mathbf{f}'\right\|^2$$

This can equivalently be written as:

$$R(\mathbf{f}') \leq R(\mathbf{f}) + \left\langle \nabla R(\mathbf{f}'), \mathbf{f}'-\mathbf{f}\right\rangle - \frac{\lambda}{2}\left\|\mathbf{f}-\mathbf{f}'\right\|^2$$

Bregman Divergence w.r.t. function R:

$$\Delta_R(\mathbf{f}'|\mathbf{f}) = R(\mathbf{f}') - R(\mathbf{f}) - \left\langle \nabla R(\mathbf{f}), \mathbf{f}'-\mathbf{f}\right\rangle$$

Clearly if a function R is λ strongly convex, then by definition, $\Delta_R(\mathbf{f}'|\mathbf{f}) \geq \frac{\lambda}{2}\left\|\mathbf{f}'-\mathbf{f}\right\|^2$

Algorithm : Let R be any strongly convex function. We define the mirror descent update as follows :

$$\nabla R(\hat{\mathbf{y}}'_{t+1}) = \nabla R(\hat{\mathbf{y}}_t) - \eta\nabla_t \quad , \quad \hat{\mathbf{y}}_{t+1} = \operatorname*{argmin}_{\hat{\mathbf{y}}\in\mathcal{F}} \Delta_R(\hat{\mathbf{y}}|\hat{\mathbf{y}}'_{t+1})$$

$$\text{Equivalently,} \qquad \hat{\mathbf{y}}_{t+1} = \operatorname*{argmin}_{\hat{\mathbf{y}}\in\mathcal{F}} \eta\langle\nabla_t, \hat{\mathbf{y}}\rangle + \Delta_R(\hat{\mathbf{y}}|\hat{\mathbf{y}}_t)$$

and we use $\hat{\mathbf{y}}_1 = \operatorname*{argmin}_{\hat{\mathbf{y}}\in\mathcal{F}} R(\hat{\mathbf{y}})$

Bound :

Claim 1. *Let R be any 1-strongly convex function. If we use the Mirror descent algorithm with $\eta = \sqrt{\frac{2\sup_{\mathbf{f}\in\mathcal{F}} R(\mathbf{f})}{nB^2}}$ then,*

$$\mathrm{Reg}_n \leq \sqrt{\frac{2B^2 \sup_{\mathbf{f}\in\mathcal{F}} R(\mathbf{f})}{n}}$$

Proof. Consider any $\mathbf{f}^* \in \mathcal{F}$, we have that,

$$\begin{aligned}\langle\nabla_t, \hat{\mathbf{y}}_t\rangle - \langle\nabla_t, \mathbf{f}^*\rangle &= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1} + \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^*\right\rangle \\ &= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\rangle + \left\langle \nabla_t, \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^*\right\rangle\end{aligned}$$

By the mirror descent update, $\nabla_t = \frac{1}{\eta}\left(\nabla R(\hat{\mathbf{y}}_t) - \nabla R(\hat{\mathbf{y}}'_{t+1})\right)$

$$= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\rangle + \frac{1}{\eta}\left\langle \nabla R(\hat{\mathbf{y}}_t) - \nabla R(\hat{\mathbf{y}}'_{t+1}), \hat{\mathbf{y}}'_{t+1} - \mathbf{f}^*\right\rangle$$

For any vectors a, b, c, $\langle\nabla R(a) - \nabla R(b), b - c\rangle = \Delta_R(c|a) - \Delta_R(c|b) - \Delta_R(b|a)$

$$= \left\langle \nabla_t, \hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\rangle + \frac{1}{\eta}\left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_R(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1})\right)$$

$\langle a, b\rangle \leq \|a\| \, \|b\|_* \leq \frac{\eta}{2} \|b\|_*^2 + \frac{1}{2\eta} \|a\|^2$

$$\leq \frac{\eta}{2} \|\nabla_t\|_*^2 + \frac{1}{2\eta} \left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|^2 + \frac{1}{\eta} \left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1}) - \Delta_R(\hat{\mathbf{y}}'_{t+1}|\hat{\mathbf{y}}_t)\right)$$

By strangle convexity of R, $\Delta_R(\hat{\mathbf{y}}_t|\hat{\mathbf{y}}'_{t+1}) \geq \frac{1}{2} \left\|\hat{\mathbf{y}}_t - \hat{\mathbf{y}}'_{t+1}\right\|^2$

$$\leq \frac{\eta}{2} \|\nabla_t\|_*^2 + \frac{1}{\eta} \left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\hat{\mathbf{y}}'_{t+1}|\hat{\mathbf{y}}_t)\right)$$

Summing over we have,

$$\sum_{t=1}^{n} \langle \nabla_t, \hat{\mathbf{y}}_t\rangle - \sum_{t=1}^{n} \langle \nabla_t, \mathbf{f}^*\rangle \leq \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_*^2 + \frac{1}{\eta} \sum_{t=1}^{n} \left(\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}'_{t+1})\right)$$

Replacing by projection only decreases the Bregman divergence

$$\begin{aligned}
&\leq \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_*^2 + \frac{1}{\eta} \sum_{t=1}^{n} (\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_t) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_{t+1})) \\
&\leq \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_*^2 + \frac{1}{\eta} (\Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_1) - \Delta_R(\mathbf{f}^*|\hat{\mathbf{y}}_{n+1})) \\
&\leq \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_*^2 + \frac{1}{\eta} R(\mathbf{f}^*) \\
&\leq \frac{\eta}{2} n B^2 + \frac{1}{\eta} \sup_{f \in \mathcal{F}} R(\mathbf{f}) \\
&= \sqrt{2B^2 \sup_{\mathbf{f} \in \mathcal{F}} R(\mathbf{f}) n}
\end{aligned}$$

Dividing through by n we prove the claim. □

2.1 Examples

Gradient Descent $R(\hat{\mathbf{y}}) = \frac{1}{2} \|\hat{\mathbf{y}}\|_2^2$. In this case mirror descent update coincides with that of Gradient descent and we recover the bound. Strong convexity is just Pythagorus theorem

Exponential Weights Let is consider the example of finite experts setting. In this setting we can consider R to be the negative entropy function,

$$R(\hat{\mathbf{y}}) = \sum_{i=1}^{d} \hat{\mathbf{y}}[i] \log(\hat{\mathbf{y}}[i]) - 1$$

Note that

$$D_R(\hat{\mathbf{y}}|\hat{\mathbf{y}}') = \mathrm{KL}(\hat{\mathbf{y}}\|\hat{\mathbf{y}}') = \sum_{i=1}^{d} \hat{\mathbf{y}}[i] \log\left(\frac{\hat{\mathbf{y}}[i]}{\hat{\mathbf{y}}'[i]}\right)$$

In this case, it is not too hard to check that R is strongly convex w.r.t. $\|\cdot\|_1$. Also note that $\sup_{\mathbf{f} \in \Delta_N} R(\mathbf{f}) \leq \log N$ (achieved at the uniform distribution).

ℓ_p and Schatten$_p$ norms Let us consider $\mathcal{F}$ to be unit ball under ℓ_p norm and $\mathcal{D}$ to be unit ball under dual norm. Let $p \in (1, 2]$, then one can use $R(\mathbf{f}) = \frac{1}{p-1} \|\mathbf{f}\|_p^2$ and this function is strongly convex w.r.t. ℓ_p norm. For matrices with analogous Schatten p norm, use the $R(\mathbf{f}) = \frac{1}{p-1} \|\mathbf{f}\|_{S_p}^2$.

Remark 2.1. *For ℓ_1 norm one can use $R(\mathbf{f}) = \frac{1}{p-1} \|\mathbf{f}\|_p^2$ with $p \approx \frac{\log d}{\log d - 1}$ and hence recover a bound of form $O\left(\sqrt{\frac{B^2 \log d}{n}}\right)$ where B is the bound on ℓ_∞ norm of ∇_t's.*